DOCUMENT RESUME

ED 392 823 TM 024 472

AUTHOR Bezruczko, Nikolaus; And Others
TITLE The Stability of Four Methods for Estimating Item Bias.
PUB DATE 89
NOTE 34p.; Portions of this report were presented at the Annual Meeting of the American Educational Research Association (1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Chi Square; Classification; *Estimation (Mathematics); Identification; *Item Bias; Racial Differences; *Reliability; *Research Methodology; Sample Size; Sex Differences; *Test Items
IDENTIFIERS Delta Coefficient; *Mantel Haenszel Procedure; *Rasch Model

ABSTRACT

The stability of bias estimates from J. Schueneman's chi-square method, the transformed Delta method, Rasch's one-parameter residual analysis, and the Mantel-Haenszel procedure, were compared across small and large samples for a data set of 30,000 cases. Bias values for 30 samples were estimated for each method, and means and variances of item bias were computed across all the samples, for comparisons contrasting sample size, sex, and race. The point estimates of item bias, based on 30 replications for each method, were also correlated across random samples, and classification techniques compared the results for agreement. The results showed that none of the methods consistently flagged more or fewer items as biased, though at the larger sample sizes the Mantel-Haenszel and Rasch methods were particularly sensitive at detecting item bias and in high agreement. Reliabilities of the Modified Delta method were generally lower than the others, as were the correlations between Modified Delta and the other indices. The results showed that not until the number of cases in each comparison group reached 1,000 did the reliabilities for any technique approach 0.80. (Contains 5 tables and 22 references. (Author/SLD)

The Stability of Four Methods for Estimating Item Bias

Nikolaus Bezruczko, E. Matthew Schulz,

Arthur J. Reynolds, Carole L. Perlman,

and William K. Rice

Department of Research and Evaluation

Chicago Public Schools

We gratefully acknowledge the assistance of Nambury S. Raju of the Illinois Institute of Technology and Linda K. Junker of the Chicago Public Schools. Professor Raju provided an item bias analysis program and technical assistance. Ms. Junker assisted in reporting results. Portions of this report were presented at the 1989 Annual Meeting of the American Educational Research Association.

ABSTRACT

The stability of bias estimates from Schueneman's Chi-square method, the transformed Delta method, Rasch's 1-parameter residual analysis, and the Mantel-Haenszel procedure, were compared across small and large samples for a dataset of 30,000 cases. Bias values for 30 samples were estimated for each method, and means and variances of item bias were computed across all the samples, for comparisons contrasting sample size, sex, and race. The point estimates of item bias, based on 30 replications for each method, were also correlated across random samples and classification techniques compared the results for agreement.

The results show that none of the methods consistently flagged more or fewer items as biased, though at the larger sample sizes the Mantel-Haenszel and Rasch methods were particularly sensitive at detecting item bias and in high agreement. Reliabilities of the Modified Delta method were generally lower than the others, as were the correlations between Modified Delta and the other indices. The results show that not until the number of cases in each comparison group reached 1,000 did the reliabilities for any technique approach .80.

INTRODUCTION

The study of differential item performance or item bias is as old as standardized testing. In 1905, in an early version of their intelligence test, Binet and Simon found group differences in test scores between working class children and those from higher social classes. Cultural bias remained an issue through the first half of the twentieth century and in 1951, Eells, Davis, Havighurst, Herrick, and Tyler (1951), published their classic study, "Intelligence and Cultural Differences."

In 1968, the study of bias entered the modern era when Cleary and Hilton (1968) published a study of item types unusually difficult for minority groups. Since then, research on methods of detecting bias has increased rapidly. Techniques have multiplied and computer applications have simplified the estimation of item bias. By 1981, Shepard, Camilli, and Averill (1981) identified 16 methods or variations of methods commonly used to detect item bias.

The development and refinement of methods for detecting item bias continues in contemporary research, motivated, in part, by courts and state legislatures throughout the country, many who now mandate the analysis of item bias for standardized tests. Some current contributions to this methodology are Shepard's modification of the Delta method, Raju's method for computing the area between two item

characteristic curves (Raju, 1987), and Holland and Thayer's recommendation (1986) to apply the Mantel-Haenszel procedure to the detection of group differences in item performance.

These developments, in detecting item bias, appear to favor formal statistical methods rather than judgmental methods such as expert raters. This emphasis on statistical methods is supported by empirical research (Qualls & Hoover, 1981) which shows little consistency in raters' perception of racial favoritism. Qualls and Hoover (1981), among others, for example, found teacher ratings for white and black raters of bias for items in the Iowa Tests of Basic Skills to have intraclass reliabilities of .03 and .11, respectively.

Despite the clarity and fairness promoted by objective statistical methods, the stability of item bias estimates has only recently been investigated systematically. Scheuneman (1980) appears to have expressed the first concern that estimates commonly produced by statistical bias techniques are unstable. Linn, Levine, Hastings and Wardrop (1981) came to the same conclusion in their analysis of four latent trait methods, and Ironson (1982) concluded that "we need more information on the reliability of the methods" (p. 152).

In the most recent study of item bias reliability, Hoover and Kolen (1984) found that six methods in common use for estimating bias "may

not lead to reliable decisions about bias" (p. 180). They suggested further studies with larger sample sizes and test items that are more clearly subject to an influence of bias.

The suggestions by Hoover and Kolen, and concerns by other researchers, are the bases for this study of the reliability of item bias estimates.

METHOD

Data

The data for this study come from the administration of multiple-choice test items in the Chicago Minimum Proficiency Skills Item Bank (Bezruczko & Reynolds, 1987). This item bank consists of approximately 1,000 minimum competency items from which an annual form of 63 items is assembled. The dichotomously scored items are structured into a three-subscale test (language arts, computation, problem solving) and administered to 30,000 eighth graders. The test is characterized by an alpha reliability of approximately .90, and a principal components analysis indicates that a single factor accounts for 30% of the variance (Reynolds & Bezruczko, 1989).

Items administered prior to 1986 were used in this study because a statistical method for estimating item bias was not yet in use, only bias review panels. This study is based on 46 items that were common to

the 1984 and 1985 forms of the test.

Subjects

Hoover and Kolen suggested that reliabilities of item bias indices be examined with larger sample sizes than they used in their study (200 per group). Consequently, in this study, 30 samples of sizes 600 and 2,000, respectively, were randomly drawn from a population of 54,986 students; both sexes and three racial groups were evenly distributed within each sample. Sampling without replacement was done insofar as possible, although it was necessary to do a small amount of sampling with replacement in order to obtain adequate Ns in each cell of the larger samples. All students were tested with 46 items, common to two test forms, over a two-year period.

Procedure

Based upon: (1) their use in prior studies (Hoover & Kolen, 1984), (2) their availability to the Chicago Public Schools, and (3) their wide use in the field (Hills, 1990), the following four methods of detecting item bias were compared: the transformed Delta method, Shephard's Modified Delta method, Rasch's one-parameter method, and the Mantel-Haenszel technique. Bias values were computed for each sample resulting in a data matrix of four methods by 46 items with 30 replications. The methods for estimating bias are described below.

One-parameter residual method. The Rasch method for assessing item bias in this study is described in Wright and Stone (1979) and Wright and Masters (1982). It involves computing Rasch item difficulty parameters and standard errors and focuses on the contrast between groups, i.e., comparison group "1", and comparison group "2", so that for every item two estimates of difficulty and precision, e.g., d(i,1), SE(i,l) and d(i,2), SE(i,2) are computed. The difference between item difficulties, expressed as a logit, for the respective groups, [d(i,l) - d(i,2)], is then standardized by SQRT[SE(i,l)**2 + SE(i,2)**2].

Mantel-Haenszel procedure. The Mantel-Haenszel (M-H) procedure for detecting item bias, as applied in this study, is described in Holland and Thayer (1988). The procedure involves matching respondents on the basis of their total test score. The total score may include or exclude the studied item. For each homogeneous level of achievement defined by matching, a 2 x 2 contingency table is constructed showing the relative performance of comparison and focal group students on the studied item. According to Holland and Thayer (1988), the number of such contingency tables depends on the number of score intervals one chooses to create for homogeneous matching. The ratio of the item's odds (p-value) for the focal and comparison groups, generalized across all 2 x 2 tables created for a particular item is

called "alpha" and is expected to be 1 if the item is not biased. In addition, the M-H procedure typically yields a Delta index and a Chi-square value with one degree of freedom. The Delta index is a log transformation of alpha:

$$[-2.35*\ln(\text{alpha})],$$

and represents the magnitude of item bias on a logit scale and should be symmetrically distributed around a mean of zero. According to Holland and Thayer (1988), the Chi-square value is used to test the hypothesis that the item is not biased.

Delta method. The Delta method of assessing item bias is described by Angoff and Ford (1973). Item p-values are calculated separately for each group and then transformed to Z-scores. Then the origin of the delta values (4Z + 13) are shifted to eliminate negative values. Item pairs of delta values are typically plotted on a bivariate graph. The delta index of item bias, which indicates the amount of bias for an item, Da, is the absolute value of the perpendicular distance between the coordinates of an item's delta values and the major axis of the ellipse of delta value pairs. To indicate which group the item favored, we attached a sign to the bias index, Da, by taking the difference between the item p-values for the contrast groups. The mean value of the signed index across the 30 replications for each item was

expected to be zero if delta is symmetrically distributed.

Modified Delta method. The Modified Delta method, described by Shepard, Camilli, and Williams (1985), improves on the Delta method by correcting for the influence of item-by-achievement interaction on the delta indices when contrast groups differ in achievement. The Modified Delta method involves regressing items' Delta indices, Da (absolute values), described above, on their point biserial correlations. The modified index is the absolute value of the residual, Ds. As described above for the Delta method, we attached a sign to Ds, based on differences in the item's p-value when comparing groups.

Computer applications. Rasch item bias analyses were performed with the programs MSCALE (Wright, Schulz, Congdon, and Rossner, 1987) and LINK (Schulz, 1984) on an IBM 3033 computer. All the other methods were performed on an XT-compatible PC using a FORTRAN program for estimating item bias developed by Raju.

Standardization of bias values and computation of reliability

Bias values. In order to simplify comparisons between methods, item bias values of all methods were converted to Z-scores. The Rasch Z-score is the logit difference between the item difficulties for comparison groups divided by its modeled standard error, as described above. If

there is no bias, the Rasch Z-score should have a mean of zero and standard deviation of 1, both within a set of 46 items and across the 30 replications of 46 items for each method.

The M-H Z-score was obtained by taking the square root of the M-H Chi-square value for an item and attaching the sign of Delta. If the assumptions regarding the Chi-square and Delta hold, this procedure should also yield a normal variate with mean zero and standard deviation 1, both within a set of 46 items and across the 30 replications for an item.

Since the Delta and Modified Delta procedures have no modeled standard errors, and produce no reference-distribution statistics, Z-scores were derived using the empirical standard deviation of the 30 bias values computed for each item. That is, for each item, within conditions of sample size and demographic contrast, i.e., different students in each sample, a bias value was estimated 30 times, as described above. The standard deviation of 30 estimates per item is an empirical estimate of the standard error of the bias index specific to the item, sample size, and demographic contrast under study. Dividing each bias value by the standard deviation of the obtained distribution yields a Z-score distribution. If the item is unbiased, this distribution also has a mean of zero, as well as a standard deviation of one.

Reliability. Reliability in this study assumes a true score model in which reliability is the ratio of the true score variance to observed variance (true score variance = observed variance minus error variance). We tested this assumption by computing reliability in two ways. First, we computed the root mean correlation, SQRT[Sum(R^2/N)], among all possible pairs of item bias values (n = 435 pairs), resulting from the 30 replications for each method. This involved taking the square root of the mean squared correlation in the 30 x 30 correlation matrix among replications, excluding the diagonal. This "root mean squared correlation" is an empirical estimate of the reliability that can also be estimated via assumptions of a true score model.

Second, based on the assumptions of a true score model, we computed the reliability coefficient, R', from estimates of the total and error variance of the bias indices: R' = (Total - Error)/Total. Total variance was the variance of all 46 (item) x 30 (replication) bias indices around their grand mean. Error variance was the pooled within-item variance of 30 indices per item around their item mean. We verified that R and R' were equivalent for all item-bias methods, and further, that it made no difference whether we used raw item bias indices or Z-scores to estimate reliability. The Delta index was used for the raw item bias index in the Mantel-Haenszel method.

Analyses

The goals of this study are to determine whether these methods differ greatly in the items they identify as biased and the extent to which their estimates are reliable. Consequently, to address the issue of intermethod consistency, we correlated the obtained item bias values--after computing mean estimates for the 30 replications for each method--between the respective methods. A correlation of 1.0 between the values for any two methods indicates that the respective methods order the amount of bias associated with the items identically.

These intermethod correlations were computed as follows: for each method at each contrast and sample size, we computed one index of bias per item by averaging the 30 estimates of bias that were obtained from the 30 samples described above. These mean estimates per item were then correlated across methods for each contrast and sample size. For example, in the male/female contrast at $\underline{n}$ = 1,000 per group, 30 standardized Rasch estimates of bias per item were averaged to give one, mean standardized estimate of bias per item. We then computed the correlation of these estimates with the similarly obtained mean standardized bias estimates for the M-H procedure.

Our second goal, establishing the reliability of these methods, involved computing procedures described above. In order to compare

methods, empirical intramethod reliabilities corresponding to pooled data from 30 samples were estimated across samples of unit size using formula 5.12.3 in Lord and Novick (1968), a standard procedure for using signal-to-noise ratio information to predict reliability coefficients for various sample sizes or test lengths. According to these procedures, for any given level of alpha reliability, the more sensitive method will flag more items as biased.

For these analyses, we chose a level of alpha reliability of .01, which is more conservative than, say .05, because the size of the data set in this study provides sufficient power to detect item bias and make reliable comparisons among methods at this probability level. In practice, when bias studies are being carried out on pilot data with many more items than will be used in a final test, typically involving a small sample size, one would probably want to choose a higher level of alpha, say .05. In general, given the same number of items in a pilot and final test, the reliability and validity of the final form will be higher if one chooses a higher alpha level for item bias detection during the development phase.

RESULTS

Mean test scores, standard deviations, and Ns for males, females, and members of the three racial groups are in Table 1. The total N for

large samples (those of size 2,000) exceeds the population size of 54,896 because some sampling with replacement was necessary in order to obtain equal numbers of persons in each of the three racial groups (no sampling with replacement was required to obtain the smaller samples).

Insert Table 1 about here

The mean proportion of times an item was flagged as biased by each method and intergroup comparison is given in Table 2. All methods flagged a larger proportion of items than were expected under the hypothesis of no bias ($p > .01$). The smallest proportions, ranging from .03 to .04, were detected in the Race1/Race2 contrast with small sample sizes (200 per group). The largest proportions, ranging from .23 to .29 were detected in the male/female contrast with large sample sizes (1,000 per group).

No one method consistently flagged a larger proportion of items than the other methods. The one-parameter residual method flagged the largest proportion (.29) in the male/female contrast. The Delta method flagged the largest proportion in the Race1/Race2 large sample size (.14) and the Race1/Race3 large sample size (.22). The Modified Delta method tied with the other methods in flagging the largest proportion in

the Race1/Race2 small sample size (.04) and Race1/Race3 small sample size (.07). The M-H method tied with the one-parameter residual method in flagging the largest proportion in the male/female small sample size (.08).

Insert Table 2 about here

Table 3 displays the summary statistics for each item bias method by sample group. Two results are apparent. First,

---

Insert Table 3 about here

---

although mean item values are consistent in large and small samples across bias methods, standard deviation and range values differ markedly. Small sample standard deviations for each method are approximately one-half of those in the larger sample. This indicates the relative lack of sensitivity in detecting item bias in small samples (individual group sizes of 300 or less). Second, with the possible exception of the male/female contrast, the methods appear to be similar in their sensitivity to item bias as standard deviation and range values are fairly similar in most intergroup comparisons. With the largest

individual group sizes (666 or more per group when $\underline{n}$ = 2,000), Mantel-Haenszel and Rasch methods were particularly sensitive in detecting bias.

Table 4 contains reliability estimates for the item bias statistics. The reliabilities for the different

---

Insert Table 4 about here

---

methods were comparable at each group size, although the Modified Delta reliabilities were slightly lower than the others. Only when the number of cases in each comparison reached 1,000 did the reliabilities approach .80.

Tables 3 and 4 present information that is useful for predicting the proportion of items that will be identified as biased using any chosen level of alpha. We simply standardized the critical values corresponding to a chosen level of alpha and found the area under the normal distribution outside their range. The standardization formula is:

$$\text{Mean} \pm \text{CV}$$

$$\text{SQRT}(\text{VARB}/\text{VARW})$$

in which CV = Absolute critical value for a chosen alpha

Mean = Empirical mean from Table 3

VARB = Between item variance from Table 4

VARW = Within item variance from Table 4

Take, for example, the Delta method in the sex contrast with 1,000 per group. Critical values with an alpha of .01 are -2.58 and +2.58. The empirical mean is .05 (Table 3). From Table 4, VARB is 4.65 and VARW is 1.0. Using these values, we get:

$$\frac{.05 \pm 2.58}{\text{SQRT}(4.65/1.0)} = -1.17 \text{ to } +1.22$$

In the standard normal distribution, about 23 percent of observations will be outside the range of -1.17 to +1.22. The results in Table 2 confirm that twenty-three percent of the Z-scores that were computed from the Delta method in the sex contrast, $\underline{n}$ = 1,000 per group, were outside the range of -2.58 to +2.58.

Similarly, we were able to predict all of the values in Table 2 by using the means and variances reported in Tables 3 and 4. We therefore feel confident that one can substitute 1.96 in the above

equation, and accurately predict the proportion of items that would be identified as biased using an alpha of .05.

The reliability of classifying a particular item with respect to a given critical value, such as 0, 1.96, or 2.58, can also be predicted from information in Table 4. The within-item variance in Table 4, VARW, is an estimate of the error variance of the Z-score bias statistic. Of course, VARW should be 1.0. VARW is necessarily 1.0 in the Delta and Modified Delta methods because we used the empirical within-item variance to compute the Z-scores. In the Rasch and M-H procedures, VARW is slightly, but consistently, less than 1. This means that significance tests based on either the M-H Chi-square or the Rasch Z-score will be somewhat conservative, and estimates of the bias of an individual item will be somewhat more reliable than one might expect because 1.0 is typically assumed to be the variance of a standard error. Table 5 indicates the high degree of similarity between item bias methods when the 30 replications for each method are aggregated across contrast groups.

Insert Table 5 about here

The diagonal elements of the matrices show the correlations of the item bias estimates within methods, across contrasts, when the 30 samples for each method are averaged. Thus, if any of the methods were repeated with another 30 samples, the correlations of the mean item bias indices would be expected to be similar to the values given on the diagonals of the matrices in Table 5. This is also the projected correlation when sample sizes are 30,000, 20,000, 9,000, or 6,000 (30 times the sample size for individual cells). The correlations in Table 5 range from .75 (n = 666, Race1/Race2, Delta/Mantel-Haenszel) to 1.00 (small and large samples, male/female, Mantel-Haenszel/Rasch). The Delta and the Modified Delta methods tend to have the lowest correlations with other methods. A high correlation coefficient in Table 5 means that any given item tends to be consistently identified as more or less biased than another item. When the correlation between any two methods is high, and alpha levels are comparable, we can conclude that any given item has been identified as biased about the same proportion of time, by the respective methods. For example, in the sex comparison Rasch and M-H methods have internal correlations of .99, and correlate with each other near 1.0. These methods are also equally powerful according to Table 4. Therefore, we can conclude that the methods are likely to agree in the proportion of times they identify any given item as

biased, or in their probability of identifying any given item as biased.

DISCUSSION

This study is based on a conception of item bias as a continuous variable, thus its obtained value is never expected to be exactly 0.0 or any specified value. Furthermore, we consider all items to have some bias whenever comparison and focal groups are not randomly equivalent. Therefore, we recommend accepting one's estimate of bias, for an item, as the best available estimate and using the theoretical estimate of error variance to determine the probability that the true bias, for an item, actually lies within an acceptable range (that is, between -2.58 and +2.58, or in one direction from a critical value such as 0).

Based on this perspective, there are several interesting results from this study. First, given the increased power of the Z-test at the larger sample size, there should be little surprise that more items were flagged as being biased using samples of size 1,000 per group rather than 300. A comparison of the methods, however, shows that none of the methods consistently flagged more or fewer items across the various comparisons or across the two different sample sizes.

Second, although the item bias methods studied here seem to be nearly equivalent in their reliabilities, at all sample sizes and demographic contrasts studied, the Rasch and Mantel-Haenszel methods

were consistently more reliable than the Delta and Modified Delta methods. The similarity, however, is deceptive.

An examination of the Rasch and M-H procedures shows that their performance depends on whether the contrast groups have equal or unequal achievement. When contrast groups (male/female) have relatively equal achievement, the Rasch and M-H bias indices correlate highly (near 1.0). This correlation is near the theoretical maximum, given the internal reliabilities of the methods when inferred from the diagonal correlations in Table 5. However, when the contrast groups have unequal achievement (Race1/Race2 and Race1/Race3) the correlation between Rasch and M-H indices is much less than their internal reliabilities.

Practitioners should also note that when a studied item is excluded from the total test score used for matching, M-H indices show a consistent overall bias favoring the higher scoring group. This relationship is reflected in Table 3 of the present study by the negative mean Z-score for the M-H method applied to the Race1/Race2 and Race1/Race3 contrasts. Recent research (Schulz, Perlman, Rice, and Wright, 1988) suggests that this overall bias in the M-H procedure is practically eliminated if one uses fine matching and includes the studied item in the total score for matching, as Holland and Thayer (1988)

recommend. However, in a strictly mathematical work, Zwick (1989) concludes that even when the studied item is included in the matching criterion, the M-H null hypothesis will not, in general, hold, and that the M-H procedure can produce a conclusion that favors either of the two contrast groups when they differ in achievement. Until further research is conducted concerning this relationship, one should perhaps be extremely cautious when selecting an item bias method when contrast groups differ significantly in achievement.

It is important to note that the Delta and Modified Delta procedures do not automatically yield Z-score indices for judging the statistical significance of bias. We were able to derive Z-scores for these methods in this study by calculating an empirical standard error from our 30 estimates of bias per item. In contrast, both the Rasch and M-H methods involve assumptions by which Z-score indices can be derived from a single bias analysis, and both provide indices for judging the scale magnitude of bias (logit difference or delta). Practitioners may therefore find the Rasch and M-H procedures more convenient, as well as slightly more reliable. One might, however, keep the Modified Delta method in mind when working with groups that differ in achievement, since it was expressly developed for this purpose, and the behavior of the Rasch and M-H methods with this type of contrast is not altogether understood at

the present time.

Third, the reliabilities and inter-method correlations presented in this study are considerably higher than those obtained by Hoover and Kolen (1984), although the reliabilities they found at smaller sample sizes (200-300 per comparison group) are hardly impressive. Since analyses of bias are necessarily test-specific and sample dependent, the differences between the two studies are not surprising, and suggest that practitioners exercise caution when attempting to generalize the results from specific studies. Although both studies indicate that reliability can definitely be a problem with smaller samples, the current study demonstrates that when the sample size of comparison groups is 666 or more, reliability of item bias indicators can be quite adequate. However, sample sizes that large are not always available.

Fourth, the selection of alpha is an important practical issue when applying methods of detecting bias. If the number of items in a test is not an issue, and the implications of having biased items is high, use a high alpha level. This rule is especially important if sample size is small. Conversely, when the question of bias is less urgent, during the piloting of new items, for example, when primary concern may be with item difficulty or guessing, the selection of alpha should be flexible. On the other hand, when items are in their fina orm, such as working elements

in a calibrated item bank, consider monitoring the stability of bias values across several administrations with more stringent alpha requirements, preferably with differing persons, at different times, from the target population.

Obviously, there are still questions that are left unanswered. For example, this study did not attempt to explain the sources of bias for flagged items. Examining the characteristics of items flagged consistently over methods (that is, by factor analysis or content classifications) may help to identify underlying sources of bias. Another question concerns a test whose items are known to have bias, either by construction or prior analysis, and the stability of the estimates under differing conditions. Also, what happens when the analyses are done on randomly equivalent groups rather than on a target and comparison group? Would the error variance of the bias indices be the same for biased items as for unbiased items? These are all questions whose answers should shed more light on the usefulness of statistical indicators of item bias.

In conclusion, these results show the fallibility of using statistical methods when sample sizes are small. They show, however, that even when sample sizes are small, as small as 300--sufficient for pilot studies--and one is willing to accept a large Type II error rate by setting alpha

high, say .1, these methods are still very useful. The results suggest that the evaluation of bias for any item, based on a statistical method, should be conducted within the context of an item's particular history of administration, rather than a single empirical analysis, and decisions regarding its reliability based on successive empirical reviews.

REFERENCES

Angoff, W. H., and Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Bezruczko, N., and Reynolds, A. J. (1987). Minimum Proficiency Skills Test: 1987 item pilot report. Citywide Report 87-1, Chicago: Chicago Public Schools.

Cleary, T. A., and Hilton, T. L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

Eells, K. A., Davis, R. J., Havighurst, R. J., Herrick, V. E., and Tyler, R. W. (1951). Intelligence and Cultural Differences. Chicago: University of Chicago Press.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement, 8, 5-11.

Holland, P. W., and Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Howard Wainer and Henry Braun (Eds.), Test Validity. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Hoover, H. D., and Kolen, M. J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8, 173-181.

Ironson, G. H. (1982). Use of Chi-square and latent trait approaches for detecting item bias. In R. A. Berk, (Ed.) Handbook of methods for detecting test bias (pp. 117-160). Baltimore: The Johns Hopkins University Press.

Linn, R. L., Levine, M. V., Hastings, C. N., and Wardrop, J. L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Qualls, A. and Hoover, H. D. (1981, April). Black and white teacher ratings of elementary achievement test items for potential race favoritism. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.

Raju, N. S. (1987). The area between two item characteristic curves. Psychometrika, in press.

Reynolds, A. J., and Bezruczko, N. (1989). Assessing the construct validity of a life skills competency test. Educational and Psychological Measurement, 49, 183-193.

Scheuneman, J. (1980, April). Consistency across administrations of certain indices of bias in test items. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Shepard, L. A., Camilli, G., and Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Shepard, L. A., Camilli, G., and Averill, M. (1980, April). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Schulz, E. M. (1984) LINK [A FORTRAN program for comparing paired Rasch estimates and linking tests through common items or persons]. Chicago: MESA Press.

Schulz, E. M., Perlman, C. L., Rice, W. K., and Wright, B.D. (1988). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing item bias. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Wright, B. D., and Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Wright, B. D., Schulz, E. M., Congdon, R. T., and Rossner, M. (1987). MSCALE [FORTRAN program for comparing paired Rasch estimates and linking tests through common items or persons]. Chicago: MESA Press.

Wright, B. D., and Stone, M. H. (1979). Best test design. Chicago: MESA Press.

Zwick, R. (1989). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? Research Report-89-32. Princeton, New Jersey: Educational Testing Service.

Table 1

Mean Test Scores by Comparison Group

	Population Size					
	Large			Small		
Group	N	Mean	SD	N	Mean	SD
Gender						
Male	30,000	32.1	8.7	9,000	32.1	8.7
Female	30,000	32.1	8.3	9,000	32.1	8.3
Race						
Race 1	19,980	35.7	7.4	6,000	35.7	7.4
Race 2	20,040	30.1	8.2	6,000	30.0	8.2
Race 3	19,980	30.6	8.6	6,000	30.7	8.6
Total:	60,000	32.1	8.5	18,000	32.1	8.5

Table 2

Mean Proportion of Times Items Were Flagged for Bias

By Method and Contrast

	Large Samples (N = 2,000)				Small Samples (N = 600)			
Contrast	D	MD	M-H	R	D	MD	M-H	R
Male/Female	.23	.23	.28	.29	.07	.07	.08	.08
Race 1/Race 2	.14	.12	.11	.12	.04	.04	.03	.03
Race 1/Race 3	.22	.21	.17	.21	.07	.07	.05	.07

Note: D = Delta, MD = Modified Delta, M-H = Mantel-Haenszel, R = Rasch.

Table 3

Mean Item Bias Z-Scores

By Method and Contrast

	Contrast								
	Male/Female			Race 1/Race 2			Race 1/Race 3		
Method	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Large Samples (N = 2000)									
Delta	.01	1.92	8.00	.03	1.48	6.89	.05	1.79	6.95
Mod-Delta	.05	1.81	7.85	.05	1.35	6.08	.17	1.62	7.55
M-H	.16	2.16	9.20	-.19	1.30	5.87	-.26	1.66	7.55
Rasch	-.11	2.20	9.34	.02	1.30	5.35	.12	1.79	8.27
Small Samples (N = 600)									
Delta	.03	1.07	4.14	.00	.82	3.75	-.03	.99	4.13
Mod-Delta	.09	.98	3.44	.09	.71	2.97	.25	.64	3.21
M-H	.09	1.12	4.85	-.09	.72	3.11	-.15	.89	3.98
Rasch	-.06	1.18	5.01	+.02	.75	3.11	.08	.99	4.66

Note: Mod-Delta = Modified Delta; M-H = Mantel-Haenszel

Table 4

Reliability of Item Bias Indices

Model	N	Contrast	Variance of Item Bias Indices: Within Items	Variance of Item Bias Indices: Between Items	Reliability
Delta	1000	Male/Female	1.00	4.65	.79
Mod-Delta	1000	Male/Female	1.00	4.24	.76
M-H	1000	Male/Female	.99	5.62	.82
Rasch	1000	Male/Female	.97	5.76	.83
Delta	300	Male/Female	1.00	2.10	.53
Mod-Delta	300	Male/Female	1.00	1.93	.48
M-H	300	Male/Female	.90	2.13	.58
Rasch	300	Male/Female	.94	2.31	.59
Delta	666	Race 1/Race 2	1.00	3.16	.68
Mod-Delta	666	Race 1/Race 2	1.00	2.79	.64
M-H	666	Race 1/Race 2	.91	2.57	.65
Rasch	666	Race 1/Race 2	.94	2.59	.64
Delta	200	Race 1/Race 2	1.00	1.64	.39
Mod-Delta	200	Race 1/Race 2	1.00	1.47	.32
M-H	200	Race 1/Race 2	.80	1.30	.38
Rasch	200	Race 1/Race 2	.89	1.44	.39
Delta	666	Race 1/Race 3	1.00	4.16	.76
Mod-Delta	666	Race 1/Race 3	1.00	3.59	.72
M-H	666	Race 1/Race 3	.90	3.63	.75
Rasch	666	Race 1/Race 3	.96	4.12	.77
Delta	200	Race 1/Race 3	1.00	1.95	.49
Mod-Delta	200	Race 1/Race 3	1.00	1.38	.27
M-H	200	Race 1/Race 3	.83	1.59	.48
Rasch	200	Race 1/Race 3	.97	1.95	.50

Note: Variance of item bias indices are estimated from 30 replications of each bias method on a 46-item test. N's give sizes of comparison and focal groups for each replication.

Table 5

Correlations Between Item Bias Methods

By Sample Size and Contrast

Sample Size		Male/Female				Race 1/Race 2				Race 1/Race 3			
		D	MD	M-H	R	D	MD	M-H	R	D	MD	M-H	R
200	D					.95	.93	.80	.99	.97	.96	.92	.90
	MD						.93	.92	.91		.92	.96	.84
	M-H							.95	.82			.97	.90
	R								.95				.97
300	D	.97	.99	.90	.90								
	MD		.97	.88	.88								
	M-H			.98	1.00								
	R				.98								
666	D					.98	.93	.75	.98	.99	.95	.88	.89
	MD						.98	.89	.92		.99	.94	.82
	M-H							.98	.81			.99	.89
	R								.98				.99
1000	D	.99	.98	.87	.86								
	MD		.99	.84	.83								
	M-H			.99	1.00								
	R				.99								

Note: These estimates of bias are based on 30 replications for a test of 46 items. The 30 estimates of bias, for each item, were then standardized as Z-scores with the mean value of item bias then the basis for computing the intermethod correlations displayed above. The intramethod correlations are estimated from between- and within-item variance in Table 4, and assume two randomly equivalent sets of data.